The Importance of Using Empirical Evidence in Software Engineering

ENRIQUE FERNÁNDEZ[1], OSCAR DIESTE[2], PATRICIA PESADO[3] AND RAMÓN GARCÍA-MARTÍNEZ[4]

[1] PhD Program on Computer Sc. School of Computer Sc. Universidad Nacional de La Plata.
[2] Empirical Software Eng. Group. School of Computer Sc. Universidad Politécnica de Madrid.
[3] Instituto de Investigaciones en Informática LIDI. Facultad de Informática. UNLP – CIC.
[4] Information Systems Research Group. Productive & Technologic Development Dept. Universidad Nacional de Lanús.
enriquefernandez@educ.ar, odieste@fi.upm.es, ppesado@lidi.info.unlp.edu.ar, rgarcia@unla.edu.ar.

Abstract. *Experiments that are run with few experimental subjects are often considered not to be very reliable and deemed, as a result, to be useless with a view to generating new knowledge. This belief is not, however, entirely correct. Today we have tools, such as meta-analysis, that we can use to aggregate small-scale experiments and output results that are equivalent to experiments run on large samples that are therefore reliable. The application of meta-analysis can overcome some of the obstacles that we come up against when running software engineering experiments (such as, for example, the practitioner availability problem).*

Keywords. *Meta-analysis, statistical power, reliability, replications, sample size.*

1. Introduction

Suppose that a hypothetical Dr. Smith is a university researcher working on testing techniques [1]. Recently, Dr. Smith has read about new inspection technique "A" that looks as if it might outperform other techniques, like, for example, technique "B". And so, he decides to run an empirical study to test this hypothesis. To do his, he puts out a call for final-year BSc in Software Engineering students to participate in the study. As a result of the all, he manages to recruit 16 students, and 8 are trained in the new technique and the other 8 in the pre-existing technique. During the experiment, each group applies the respective technique to the same program. He measures the number of effects detected as the response variable. Table 1 shows the results (aggregated by group).

Table 1. Results of the experimental study by Dr. Smith

Technique A	Technique B
Means (Y_e) = 12.000 defects	Means (Y_c) = 11.125 defects
Standard Deviation (S_e) = 2.673	Standard Deviation (S_c) = 2.800

Based on these values, Dr. Smith runs a hypothesis test (a t-test assuming variances to be equal) with $\alpha = 0.05$. This test returns a p-value of 0.53. Therefore, technique A cannot be said to perform better than B.
Although the results are not promising, Dr. Smith decides to go ahead with their publication in the hope that the experiment will be replicated and the aggregation of data will better explain the comparison between A and B. Dr. Smith submits the paper and, at the end of the review process, receives the assessment shown in Figure 1.

Originality:	*Neutral*
Importance:	*Strong Reject*
Overall:	*Reject*
Detailed comments:	*Your paper is interesting but has two major pitfalls. First, it was developed with very few experimental subjects (which, on top of this, are not practitioners). Second, the study results are not significant, meaning that it provides no useful information.*

Fig. 1. Results of the paper review process

The above example, albeit fictitious, is representative of many real pieces of empirical software engineering (ESE) research. On the one hand, many researchers interpret hypothesis testing too restrictively (and wrong in many cases, as will be seen later), focusing on whether or not the results are significant (level $\alpha = 0.05$). On the other hand, there is a tendency not to take experimental studies that were built with students as evidence, as this research is not considered to be extrapolable to real-world environments. However, there is a shortage of subjects (be they practitioners or students) that are willing to participate in experimental studies [2][3][4]. Additionally, the more subjects an experiment has, the more costly it will be in terms of workload, infrastructure, among others, and this can discourage researchers. On the other hand, the cost of experiments run with fewer subjects is likely to be more affordable. These factors clearly limit SE researchers' prospects of being able to generate new empirically validated knowledge.
Fortunately, there are some alternatives for exploiting the results of small-scale studies. In this paper we will focus on one: meta-analysis. Essentially, meta-analysis is a statistical technique for aggregating more than one study, thereby increasing the number of experimental subjects involved in the hypothesis testing and outputting more reliable results. In our research we have analyzed whether meta-analysis could be applied in ESE to combine the

results of several small-scale experiments, with the aim of increasing the power of experiments with small samples.
We will proceed as follows. Section 2 will describe how sample size affects hypothesis testing. In Section 3 we will outline how to use meta-analysis to combine the results of more than one small study and thus increase their power. Section 4 presents a set of problems related to meta-analysys. Finally, Section 5 will discuss whether meta-analysis is reliable when applied to ESE.

2. State of the Art

Any statistical test is subject to two types of errors: α, or type I error, and β, or type II error [5]. These errors occur due to the uncertainty associated with estimating population parameters (means and standard deviation) from a sample of the population. Remember that an experiment observes what happens in a sample (the subjects that tested the techniques) to estimate what happens in a population (the reality of the tested techniques). As Table 2 shows, α is the error associated with the alternative hypothesis (H_1: there is a difference between tested techniques) being accepted when the null hypothesis (H_0: there is no difference between the tested techniques) holds for the population, and β is the likelihood associated with the opposite event.

Table 2. Decision of the statistical test

	H_0	H_1
H_0	Correct decision (1-α)	β (Tipe II error)
H_1	α (Tipe I error)	Correct decision (1-β)

It is more dangerous for an experiment to lead to the belief that there actually is a difference between two tested techniques when there really is none (error α) than to believe that there is no difference (because none is observed in the experimental sample) when there really is (error β). Therefore, the value of α is set at extremely low values, such as 0.1, 0.05 or even 0.01 (10%, 5% and 1%, respectively).
Unfortunately, α and β are not independent: according to statistical theory, a hypothesis test is characterized by five factors [6]: α, β, the mean difference *d*, the level of variation of the response variable *s* (measured as the variance or standard deviation) and the number of experimental subjects, or, to be more precise, sample size, *n*. Equation 1 shows the relationship between these factors, where *z* represents the typified normal distribution:

$$z_{1-\beta} = \sqrt{\frac{n}{2}}\frac{d}{S} - z_{1-\alpha} \qquad (1)$$

These five factors form a closed system. This means that an increase or decrease in any one of the factors leads to increases or decreases in the others. In practice, the factor that really is affected is *n*, as type I (α) and type II (β) errors are set beforehand, and both *d* and *s* are circumscribed by the experimental context and cannot therefore be manipulated at liberty by the researcher. This is perhaps the most important, albeit not the only, reason why experiments are required to have a large number of experimental subjects. When the number of experimental subjects is small and α is set at 0.05, β returns very high values.

Let us go back to the example of Dr. Smith's experiment. Applying Equation 1 we get β = 0.83, that is, the test will detect significant differences 17% of the time, whereas it will fail to do so in 83% of the cases, even though they possibly do exist in the population/reality. The influence of the number of subjects on type II error is even clearer if we look at how β decreases as more experimental subjects join, all other factors being equal. It is usual practice to use the term *reliability* instead of α to refer to type I error and *statistical power* instead of β to refer to type II error. Reliability is calculated as 1 - α and power as 1 - β. For an experiment to be considered reliable, it is usual to set type I error at α = 0.05 (that is, a reliability of 0.95 or 95%) and type II error at β = 0.2 (that is, a power of 0.8 or 80%). As Figure 2 shows, Dr. Smith would have needed a total of 120 subjects (60 in each group) for her experiment to be considered reliable. Fortunately, there are several strategies designed to overcome the problems of low power caused by the use of experiments that have few experimental subjects. In this paper, we will look at meta-analysis.

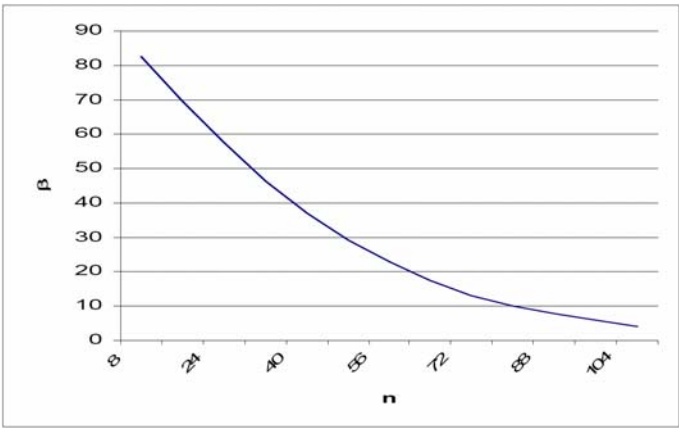

Fig. 2. Decrease in type II error against the increase in n

3. Taking Advantage of Experiments with Few Subjects

Meta-analysis is a statistical technique for combining the results of more than one experiment developed previously to achieve a greater statistical power

than any of the individual experiments on their own. Although usually associated with medicine, the term meta-analysis as it is now known was developed in psychology.

In many cases of psychology, the treatments studied have very small effects on experimental subjects, meaning, as illustrated in Figure 2 [7], that experiments need a very large sample size (usual guidelines are around 150 [8]). In many cases, however, not that many subjects are available for experiments and studies reporting insignificant effects predominate over others that do detect significant effects, as studies of low statistical power accumulate. This was the way things were in psychotherapy, the specialized field with which Dr. G.V. Glass [9], creator of meta-analysis as we know it today, was concerned. Using an argument very similar to the one brandished in ESE today (small studies are useless), psychotherapy was judged to be ineffective. Dr. Glass, who did not agree with this interpretation, took a different road to demonstrate his belief: instead of excluding studies (on the grounds of their size or statistical significance), he tried to consider as many studies as possible upon which to base his findings. Looking back, the hardest thing was to find a way of aligning the wide range of metrics used in the different replications to measure the response variables. The solution was to come up with what is today the well-known concept of effect size, briefly mentioned in Section 2. Effect size is a non-scalar measure calculated as the difference between the treatment means divided by the pooled standard deviation.

After calculating the effect size of each experiment, all Glass had to do was average the results of the individual experiments to arrive at a *global effect* using a procedure dating back to the mid-19th century[10]. This value represents the effect that, theoretically, a single experiment having a greater sample size and, consequently, a smaller type II error than any of the original experiments would have achieved. This way he demonstrated that psychotherapy was indeed effective. The parallelisms with ESE, in respect of the potential contribution of small studies, are evident. For example, suppose that Dr. Smith published her paper on her laboratory web site. Later Dr. Thomas visited the web site, found the experiment interesting and decided to replicate it. In this case, Dr. Thomas managed to recruit no more than eight advanced MSc in Software Engineering students, four of which he assigned to each of the experimental groups. Results are shown in Table 3.

Table 3. Results of Dr. Thomas' experimental study

Technique A	**Technique B**
Means (Y_e) = 13.000 defects Standard Deviation (S_e) = 1.800	Means (Y_c) = 12.000 defects Standard Deviation (S_c) = 1.700

Dr. Thomas ran a t-test on these results (assuming variances to be equal at $\alpha = 0.05$) and also found insignificant differences (pvalue = 0.57). What would happen if these two studies were combined using meta-analysis to achieve a new result? Would the differences be significant? Would the test be more powerful? In response to these questions, the sample size is still not big

enough to return significant results (there are only 12 subjects per technique). Figure 3 (showing the statistical power of the meta-analysis for a population with an effect size 2 of 0.5 and α=0.05) indicates that about 70 experimental subjects would be necessary for a meta-analysis to achieve what is usually considered as a discriminative statistical power ($1-\beta = 0.8$).
The statistical power, however, has improved in part. Whereas Dr. Smith's and Dr. Thomas' tests had a power of 0.17 and 0.11, respectively, the meta-analysis achieved a power of 0.13. Note that if it had been possible to use 8 + 4 = 12 subjects per group in a single experiment, it would have been possible to achieve a power of 0.22. Using meta-analysis it is possible to gradually increase the statistical power as more experiments are added. This way experiments with a small sample size can supplement each other. The more experiments (no matter how small the number of subjects per experiment is) that are aggregated using meta-analysis, the more powerful the results and, consequently, the greater the possibility of detecting false-negatives will be.

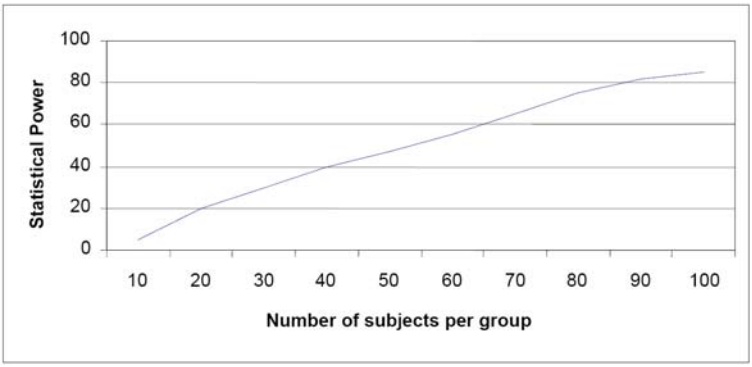

Fig. 3. Increase of the statistical power in a meta-analysis

Suppose that there are three more replications of Dr. Smith's research, whose results are shown in Table 4 (note that they all return insignificant results).

Table 4. Results of Identified Studies (Replications)

Study	Ne	Me	Se	Nc	Mc	Sc	p-value	Power
3	9.00	11.00	1.80	9.00	10.10	1.70	0.30	17.58
4	10.00	10.00	1.40	10.00	9.10	1.50	0.20	20.34
5	12.00	9.00	1.60	12.00	8.10	1.80	0.20	24.63

Figure 4 charts how the power of the meta-analysis increases as more of these studies are added. In this example, even though the test fails to achieve the desired power level of 80%, it does, in any case, manage to output significant differences at a power of almost 57% (which is much greater than the best experiment separately, estimated at 24%).

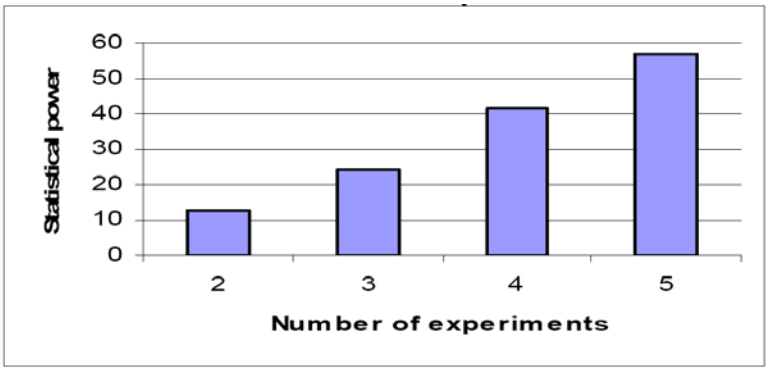

Fig. 4. Increase in the statistical power by accumulating small-scale replications

This is noteworthy, as the example was designed based on the fact that the inspection technique efficiency actually IS different.
So far we have given an example of how meta-analysis can be used to take advantage of studies with small sample sizes that, separately, return results that are insignificant but, together, could provide valuable evidence.

4. Theoretical Model Limitations

Although the functions of estimating the power of a meta-analysis to estimate accurately the statistical power of a meta-analysis when the set of experiments that are part of the aggregation process are homogeneous (the differences between the results of different experiments are minimal), in our view, this function has shortcomings to be applied in the current context of the SE. That problem is that such studies are small, variations in the results, in general, are great influences of experimental error. When this happens the power of meta-analysis tends to decline as indicated obliquely by Hedges and Olkin [11]. We have analyzed this aspect using a Monte Carlo test, which simulate the results of studies within the same population, but without forcing homogeneity among the groups to add, varying the number of subjects for experiments between 4 and 20 and combining 2 to 10 experiments for meta-analysis.
This study allowed us to determine when it has no homogeneity, that for low effect sizes (0.2) is almost impossible to get the test showed significant differences working with 200 subjects per group (simulated maximum sample size) due to the low power statistical effects that for effect sizes medium and high (0.5 and 0.8) heterogeneity is not determinant, as it has been able to achieve good statistical power with lots of subjects no too high (about 80 to 30 subjects per group, respectively). Figure 5 shows a comparison of the estimated power for each of the values of typical effect.

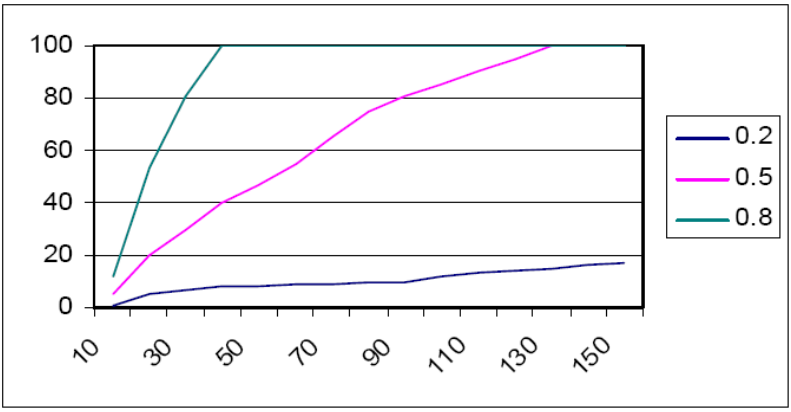

Fig. 5. Simulated power for a meta-analysis

5. Conclusions

In this paper we have shown that there are options open to researchers to generate pieces of empirical SE knowledge more efficiently than they do today. We have shown that meta-analysis is able to increase the power of experiments, enabling a set of small studies that individually do not return statistically significant differences to do so, if taken together. This way we can solve some of the problems related to the accumulation of a sizeable number of experimental subjects by a single researcher, as we can put together a large-scale experiment by meta-analyzing replications of small studies.
In summary, we can say that:

[1] It is worthwhile running experiments even if they do not have many experimental subjects, as they can be combined to form a larger scale study;

[2] It is worthwhile publishing studies even if they do not return significant results, as this can be very often due to the low power of the statistical method.

[3] If this strategy were applied to really important technologies in SE (i.e.: UML or partition of equivalence), the combined effort of the investigators would allow to decide whether these technologies are really useful or not, providing the necessary foundation to become software development in an engineering process.

Acknowledgments

This research has been partially funded by grant UNLa-SCyT-33081 of the National University of Lanus (Argentine) and by grants TIN2008-00555 and HD2008-00046 of the Spanish Ministry of Science and Innovation (Spain).

References

1. Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sörumgård, S., Zelkowitz, M.; 1996; *The empirical investigation of perspective-based reading*, International Journal on Empirical Software Engineering, Vol. 1, No. 2, 133-164.
2. N. Juristo, A. Moreno (2001). Basics of Software Engineering Experimentation, Kluwer Academic Publishers.
3. Tonella P., Torchiano M., Du Bois B., Systä T. (2007). *Empirical studies in reverse engineering: state of the art and future trends*; Empir Software Eng 12:551-571.
4. Dyba, T., Aricholm, E.; Sjoberg, D.; Hannay J.; Shull, F. (2007). Are two heads better than one? On the effectiveness of pair programming. IEEE Software, 12-15.
5. Everitt, B. (2003). *The Cambridge Dictionary of Statistics*, CUP.
6. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* (2nd ed.).
7. Gurevitch, J. and Hedges, L. (20019. *Meta-analysis: Combining results of independent experiments.* Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), 347-369. Oxford University Press, Oxford.
8. Fisher RA (1925). *Statistical Methods for Research Workers* (first ed.). Edinburgh: Oliver & Boyd.
9. Glass, G. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher 5: 3-8.
10. Cochrane (2008). Curso Avanzado de Revisiones Sistemáticas; www.cochrane.es/?q=es/ node/198.
11. Hedges, L.; Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press.